Ozone measurements monitoring using data-based approach

Fouzi Harrou[a,*], *Farid Kadri*[b], *Sofiane Khadraoui*[c], *Ying Sun*[a]

[a] *CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*
[b] *PIMM Laboratory, UMR CNRS 800, Arts et Métiers ParisTech, Paris, France*
[c] *University of Sharjah, Department of Electrical and Computer Engineering, Sharjah, United Arab Emirates*

A B S T R A C T

The complexity of ozone (O_3) formation mechanisms in the troposphere makes the fast and accurate modeling of ozone very challenging. In the absence of a process model, principal component analysis (PCA) has been extensively used as a data-based monitoring technique for highly correlated process variables; however, conventional PCA-based detection indices often fail to detect small or moderate anomalies. In this work, we propose an innovative method for detecting small anomalies in highly correlated multivariate data. The developed method combines the multivariate exponentially weighted moving average (MEWMA) monitoring scheme with PCA modeling in order to enhance anomaly detection performance. Such a choice is mainly motivated by the greater ability of the MEWMA monitoring scheme to detect small changes in the process mean. The proposed PCA-based MEWMA monitoring scheme is successfully applied to ozone measurements data collected from Upper Normandy region, France, via the network of air quality monitoring stations. The detection results of the proposed method are compared to that declared by Air Normand air monitoring association.

Keywords:
Anomaly detection
MEWMA statistic
MSPC
Principal components analysis
Ozone pollution
Data-driven strategy

1. Introduction

Atmospheric pollution is one of the most serious problems confronting our modern world. The impact of atmospheric pollution on human health is now forefront of population concerns (Moshammer, 2010). Numerous epidemiological studies highlight the influence on the health of certain chemical compounds such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) or dust particle in the air (Moshammer, 2010). The influence of this pollution is noticeable on sensitive populations such as asthmatics, children, and elderly. Currently, among the monitored compounds, ozone is one of the greatest concern. Ozone is one of the most important photochemical oxidant that exerts adverse effects on human health as well as damages ecosystems, agricultural crops and materials at certain concentration levels (Nawahda, 2016; Sillman, 2003; Chiogna and Pauli, 2011). France, like most European countries, has often known during the last summer seasons (2003 especially) episodes of ozone pollution, affecting a large part of the territory. The detection of abnormal pollution in the measured concentrations of these compounds is therefore an important issue for health.

The acceptable concentrations of these pollutants, harmful for human health and the environment, are defined by European standards. Air quality monitoring networks have the following main missions: the measurement network management (recording of pollutant concentrations and a range of meteorological parameters related to pollution events) and the diffusion of data for permanent information of population and public authorities in reference to norms. The objective of

* *Corresponding author*. Tel.: +966 546326240; fax: +974 012 8080602.
E-mail addresses: fauzi.harrou@kaust.edu.sa (F. Harrou), farid.kadri@ensam.eu (F. Kadri), sofiane.khadraoui@qatar.tamu.edu (S. Khadraoui), ying.sun@kaust.edu.sa (Y. Sun).

this work is to propose a statistical detection method able to detect abnormal ozone measurements caused by air pollution or any incoherence between the different network sensors or sensor dysfunction. The complexity of ozone (O_3) formation mechanisms in the troposphere (Seinfeld and Pandis, 2006), the complexity of meteorological conditions in urban areas and the uncertainty in the measurements of all the parameters involved, make the fast and accurate modeling of O_3 very challenging. As an alternative, implicit modeling approaches, which are data-based techniques (like principal component analysis), are particularly well adapted to reveal linear relationships among the process variables without formulating them explicitly. To overcome this difficulty, the principal component analysis (PCA) (a basic method in the framework of multivariate analysis techniques) can be used because they need no prior knowledge about the process model (Yin et al., 2014). PCA is one of the most popular multivariate statistical technique used in extracting information from data and is widely used by scientists and engineers in various disciplines, such as in face recognition, data compression, image analysis, visualization, as well as in anomaly detection (Qin, 2003; Herve and Lynne, 2010; Yin et al., 2014). In the absence of a process model, principal component analysis (PCA) has been successfully used as a data-based anomaly detection technique for highly correlated process variables (Qin, 2003). Due to its simplicity and efficiency in processing huge amount of process data, it is recognized as a powerful tool of statistical process monitoring (Qin, 2012; Khan et al., 2015). PCA and its extensions has been successfully applied in a wide range of applications, such as in chemical processes (Banimostafa et al., 2012), water treatment (George et al., 2009) and hospital management (Harrou et al., 2015).

Generally, in PCA based process monitoring, PCA develop a reference model using the normal data collected from the normal process. The new process behavior can thus be compared with the predefined one by the monitoring system to ensure whether it remain under normal operating conditions or not. When anomaly occurs, the process moves out of the normal operation regions indicating that the change in the process behaviors has occurred. Typically, Hotelling T^2 statistic (Hotelling, 1933) and the sum of squared residuals SPE (Box, 1954) which is also known as the Q statistic (Romagnoli and Palazoglu, 2006) are used in PCA-based monitoring to elucidate the pattern variations in the model and residual subspaces, respectively. The T^2 statistic is defined by the Mahalanobis distance whereas the Q statistic is defined by the Euclidean distance to avoid ill-conditioning due to small eigenvalues (Geladi and Kowalski, 1986; Kourti and MacGregor, 1995; MacGregor and Kourti, 1995; Qin, 2003; Chen et al., 2004). In other words, the T^2 statistic is a measure of the variation in the PCA model and the Q statistic is a measure of the amount of variation not captured by the PCA model. The main disadvantage of using PCs in process monitoring is the lack of physical interpretation (Ranger and Alt, 1996; Kourti and MacGregor, 1996). In addition, in the previous study, Romagnoli and Palazoglu (2006) have shown that the T^2 statistic can result in false negatives (missed detection) due to the latent space sometimes being insensitive to moderate process upsets, which is because each latent variable is a combination of all process variables. Additionally, the disadvantage of T^2 statistic is that anomalies in the process mean that are orthogonal to the first PCs cannot be detected by using the T^2 (Mastrangelo et al., 1996). The Q statistic, however, is more sensitive to additive anomalies than the T^2 statistic because additive anomalies propagate to the model error. However, the Q statistic can better detect changes in the correlations between the process variables than T^2 (Harkat et al., 2006), and is also more sensitive than T^2 to modeling errors (Harkat et al., 2006). Nevertheless, the major disadvantage of the conventional PCA-based detection indices, is that use only the information enclosed about the process in the last observation and they ignore information given by the sequence of all observations. Consequently, this make these detection indices relatively insensitive to small changes in the process variables (Montgomery, 2005). These shortcomings of the T^2 and Q statistics motivate the use of other alternatives in order to mitigate these disadvantages. To overcome the previous shortcomings, an alternative approach is proposed in this paper, in which PCA is used as a modeling framework in a model-based anomaly detection method. In this approach, PCA is used to express a process data matrix as the sum of approximate and residual matrices. After a model is obtained using PCA, various methods for anomaly detection can be applied, such as the multivariate exponentially weighted moving average (MEWMA) monitoring scheme, which is utilized in this work to improve anomaly detection. Therefore, the main contribution of the paper is to exploit the greater ability of the MEWMA monitoring scheme to detect small shifts in the process mean for improved anomaly detection of conventional PCA. More specifically, this paper proposes PCA based-MEWMA anomaly detection methodology for detecting abnormal ozone measurements.

The rest of this paper is organized as follows. Section 2 provides a brief overview of ground-level ozone (i.e., tropospheric ozone) pollution. The used data sets and study site are described in Section 3. Then, PCA and a description of how it can be used in anomaly detection is presented in Section 4. Next, the multivariate EWMA which is commonly used in quality control is described in Section 5. Then, the proposed PCA-based MEWMA anomaly detection approach, that integrates PCA modeling and MEWMA monitoring scheme, is presented in Section 6. In Section 7, we present the application of the PCA-based MEWMA anomaly detection approach to detect abnormal ozone measurements of an air quality monitoring network in Upper Normandy, France. Conclusions and future works are finally presented in the last section.

2. Ozone pollution

Generally, two types of ozone are distinguished: (1) Stratospheric or good ozone, present at around 13–30 km of altitude, is a natural filter that protects life on earth from the harmful (ultraviolet) rays of the sun (Sillman, 2003). The ozone hole is a partial disappearance of this filter, linked to the ozone destroying effects of certain pollutants emitted into the troposphere and that move slowly into the stratosphere. (2) Tropospheric ozone or ground-level ozone, present in the air we breathe, is bad: it causes eye irritation, bronchial, and can cause respiratory problems, especially among vulnerable persons (children, elderly) or asthma. The tropospheric ozone (O_3) is a pollutant that has attracted growing interest in recent years (Vlachokostas et al., 2010; Detournay et al., 2007). Unlike other pollutants, ozone is not directly emitted to the atmosphere. It is a pollutant called secondary formed as a result of complex chemical reactions involving two large families of pollutants known as primary: volatile organic compounds (VOC) and industrial emissions release a family of nitrogen

oxides (NOx) (Brulfert et al., 2007). It is formed gradually under the action of solar radiation (NOx and VOC combine chemically with oxygen to form ozone during sunny) and ozone important peaks can be seen in the summer. High levels of ozone are usually formed in the heat of the afternoon and early evening, dissipating during the cooler nights. The tropospheric ozone is a pollutant that must be monitored. Ozone, O_3, is produced by the reaction represented by the following equation:

$$NO_2 + O_2 \longrightarrow NO + O_3$$

where NO_2 is the nitrogen dioxide, NO is nitrogen monoxide and O_2 is the oxygen. The nitrogen oxides (NO_2) result from the combination of oxygen (O_2) with nitric oxide (NO) induced by human activities (combustion of hydrocarbons, for transportation or heating...) and volatile organic compounds (VOCs) mainly coming from industries. Solar radiation of wavelengths less than 430 nm are capable of dissociating NO_2 into a molecule of nitric oxide (NO) and oxygen (O). This last is combined with the oxygen to form the molecule of ozone (O_3).

This reaction provides two essential information: (i) ozone photochemical pollutant is formed only during daylight hours under appropriate conditions, but is destroyed throughout the day and night. Ozone concentrations are higher on hot, sunny, calm days. Generally, ozone concentration is highest in the rural sites than the urban sites. Higher concentrations in rural areas can be result from nitrogen oxides and volatile organic compounds being transported from upwind urban or industrial areas, by natural ozone being transported to ground-level from the upper atmosphere, or from natural volatile organic compounds emitted from vegetation (Due nas et al., 2004; Proyou et al., 1991). (ii) At night, ozone produced in the light of day (due to direct solar radiation), disappears. This is due to the destruction of ozone by nitric oxide, which is emitted by vehicles. Nitric oxide can remove ozone by reacting with it to form nitrogen dioxide ($3NO + O_3 \rightarrow 3NO_2$). Ironically, the concentrations of nitric oxide are very low in most rural areas to completely destroy ozone, so ozone remains in the atmosphere for a longer period. Ozone levels tend to be higher in rural areas where there are less local emissions of nitrogen dioxides to destroy any ozone that has formed in the atmosphere (Brankov et al., 2003).

2.1. *Diurnal variation of ground-level ozone*

Diurnal variations of ozone concentrations follow a typical cycle, with a minimum in late night and a maximum around mid afternoon (Chen et al., 2015), as shown in Fig. 1. This figure shows the measurements of seven different stations (located in the same network) for the same day. The seven curves have a daily behavior very similar. The ozone concentration begins to increase just after sunrise, and attains its maximum level in the afternoon due to photochemical production of O_3 mainly from oxidation of natural and anthropogenic hydrocarbons, carbon monoxide (CO), and methane (CH_4) by hydroxyl (OH) radical in the presence of a sufficient amount of NOx.

2.2. *Anomalies in ozone measurements*

Two types of anomalies in ozone measurements (atypical ozone peaks) can be distinguished: true and false anomalies. True anomalies correspond to peaks in the ozone levels due to the production of photochemical ozone. The formation of a true peak of ozone requires certain conditions, such as sunny days under stagnant and humid air conditions, high humidity and high temperatures to promote the formation of ozone, and low wind speeds to accumulate high pollution levels. These peaks are usually large with a duration of several hours (due to long reaction times needed for a gradual formation of the photochemical ozone). Therefore, this type of anomalies usually exhibit bell shaped curves. Furthermore, false anomaly are usually observed outside the summer period, where the ozone concentration abruptly increases with very high ozone concentrations (to be in the range of 150–600 $\mu g/m^3$) for short periods of time (around 1 h). These abnormal measurements are sharply pointed, which are different from those observed in the case of photochemical ozone. The presence of this type of anomalies can be due to different phenomena: (a) malfunctioning sensor(s), (b) transported ozone produced elsewhere in the region, (c) transported ozone produced elsewhere in the region, and others (Zdanevitch, 2001) (Fig. 2).

3. Air quality monitoring in French using network of measurement stations

Pollution of the lower atmosphere by ozone is a growing problem in industrialized countries. In France, the law on air quality and rational use of energy (LAURE, law n° 96-1236, 30th December 1996) provides a set of measures to guarantee for citizens the best air quality. Hence the fight against air pollution becomes a priority. Today, according to this law, all cities in France, with more than 100 000 inhabitants, have an air quality monitoring network. Actually in France, we have 40 networks where each of them is managed by a local association. Fourteen air quality monitoring associations (AASQA) have been created and approved by the ministry of environment to monitor air quality in France. Atmo federation groups all these 40 approved associations. These associations measure, collect, monitor and observe air quality. AASQA continuously monitor the presence in the ambient air of 13 pollutants regulated by European directives and national legislation. Ozone (O_3) forms part of the pollutants which are measured by monitoring associations, because it can cause a number of respiratory health effects. Monitoring networks for air quality generally consist of several measuring stations spread over the geographical area concerned. When the air pollutant concentration exceed a certain threshold (defined by decree in air quality regulations) or there is a risk to exceed it, the association is in charge to inform general public with information on the measured values and to give advices/recommendations for the exposed populations.

The heat wave of summer 2003 in France, was linked with an exceptional ozone pollution, that affected the whole European community. These levels were specially high and related to the weather conditions and exceptional temperatures. The consequences of this heat wave demonstrated the importance to dispose of reliable warning systems for detection of unexpected pollution and unforeseeable events. Considerable efforts have been deployed (and still are) to equip AASQA by descriptive models of ozone dispersion. However, we can notice that these so-called deterministic models are sometimes far from reality. Hence, it is important to propose new optimal descriptive models and statistical methodology for the detection of peak ozone levels. This will be the principal objective of this study. In the next subsection the ozone data set used in this study will be briefly described.

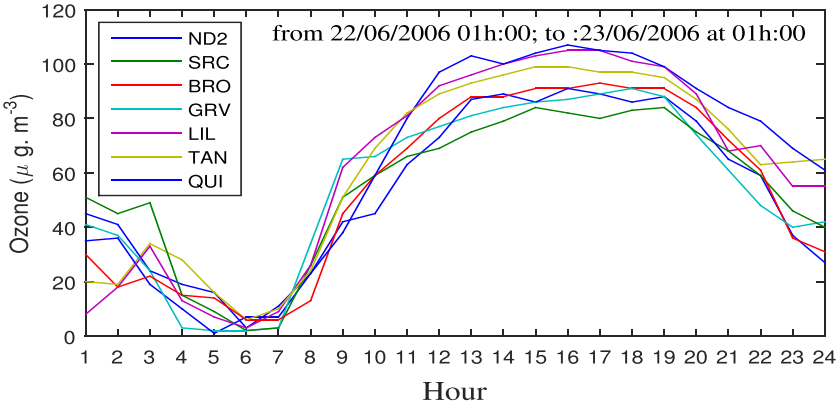

Fig. 1 – Example of daily ozone concentrations.

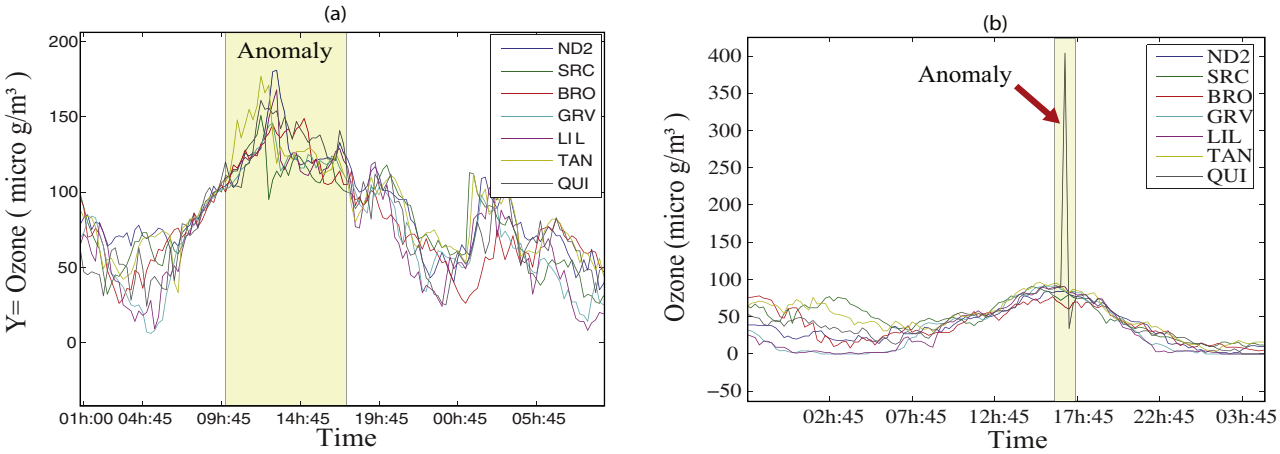

Fig. 2 – Types of ozone anomalies: (a) true anomaly and (b) false anomaly.

3.1. *Data sets and study region*

In this study, the Upper Normandy region was selected for data collection. Upper Normandy is located at northwest of Paris, near the south side of Manche sea and is one of the most highly industrialized areas in France. This city, like most large European cities, faces air pollution problems. The association Air Normand is the official association responsible for monitoring air quality over Upper Normandy region, and providing with information on the results.

Generally, there are different types of air quality monitoring stations: local, urban, rural and industrial. The local stations, directly exposed to industrial locations or positioned close to traffic, convey the concentration of pollutants emanating from an identified source. The urban stations measure the ambient air pollution to which the majority of the population is exposed. Finally, the rural stations are representative of the levels observed in the sparsely populated areas and enable the long distance consequences to be assessed. Each station consists of a set of sensors, dedicated to the acquisition of pollutants (ozone O_3, nitrogen oxides NO, sulfur dioxide SO_2,...). In order to measure and control tropospheric ozone pollution the Air Normand association consists of seven stations placed in industrial, peri-urban and urban sites, across the region. Ozone concentrations have been measured every 15 min by Air Normand network. Fig. 3 shows a map of France and the location of study sites (Champagne-Ardenne and Upper normandy).

The aim of this study is to apply the proposed PCA-based MEWMA anomaly detection algorithm in order to detect

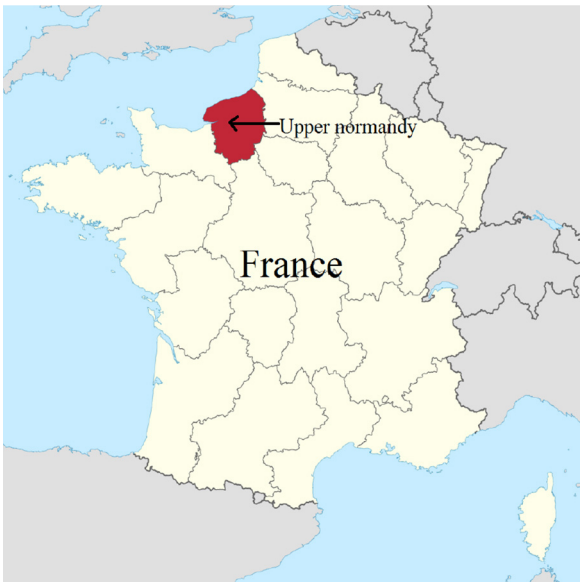

Fig. 3 – Location of study sites in France (http://education.francetv.fr/CartesInteractives).

abnormal measurements of ozone, both of anthropogenic origin (pollution peaks caused by human activity) or the result of dysfunction of sensors (anomalies, interference,...). A brief introduction to the principles of PCA, and how it can be used in anomaly detection is presented next.

4. Principal component analysis (PCA)

PCA is a linear dimensionality reduction modeling method, which can be helpful when handling data with a high degree of cross correlation among the variables. The main idea behind PCA is briefly introduced in this section, and more details can be found in Patton and Chen (1991) and MacGregor and Kourti (1995).

4.1. PCA modeling

Let us consider the following raw data matrix $\mathbf{X} = \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T\right]^T \in R^{n\times m}$ consisting of n observations and m correlated variables. The data are collected when the monitored process is under normal operating condition so that the PCA's model that will be built represents a reference of the normal process behavior. Before computing the PCA model, the raw data matrix $\mathbf{X}$ is usually pre-processed by scaling every variable to have zero mean and unit variance. This is because variables are measured with various means and standard deviations in different units. This pre-processing step puts all variables on an equal basis for analysis (Ralston et al., 2001). Let $\mathbf{X}_s$ denotes the autoscaled matrix of $\mathbf{X}$. By using singular value decomposition (SVD), PCA transforms the data matrix $\mathbf{X}_s$ into a new matrix $\mathbf{T} = [t_1\ t_2 \cdots t_m] \in R^{n\times m}$ of uncorrelated variable called score or principal components (PCs). Indeed, PCs are just mathematical constructs chosen to represent the variance as efficiently as possible, even if their physical meaning is obscure. Each principal component is a linear combination of the original variables, so that $\mathbf{T}$ is obtained from $\mathbf{X}_s$ by an orthogonal transformations (rotations) designed by $\mathbf{P} = [p_1\ p_2 \cdots p_m] \in R^{m\times m}$ which is given as follows:

$$\mathbf{T} = \mathbf{X}_s\mathbf{P} \quad \text{and} \quad \mathbf{X}_s = \mathbf{T}\mathbf{P}^T = \sum_{i=1}^{m} t_i p_i^T, \tag{1}$$

where the column vectors $p_i \in R^m$ of the matrix $\mathbf{P} \in R^{m\times m}$ (also known as the loading vectors) are formed by the eigenvectors associated with the covariance matrix of $\mathbf{X}_s$, i.e., Σ. The covariance matrix, Σ, is defined as follows:

$$\Sigma = \frac{1}{n-1}\mathbf{X}_s^T\mathbf{X}_s = P\Lambda P^T \quad \text{with} \quad PP^T = P^TP = I_n, \tag{2}$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_m)$ is a diagonal matrix containing the eigenvalues in a decreasing order ($\lambda_1 > \lambda_2 > \cdots > \lambda_m$), I_n is the identity matrix, and the ith eigenvalue equals the square of the ith singular value (i.e., $\lambda_i = \sigma^2$) (Jackson and Mudholkar, 1979).

Note that the PCA model results in the same number of principal components as the number of originals variables (m). In the case of collinear process variables, however, a smaller number of principal components (l) are needed to capture most of the variations in the data. Often, a small subset of the principal components (corresponding to the largest eigenvalues) can extract most of the important information in a data set, and thus simplify its analysis. The first PC indicates the direction of largest variation in data, the second PC indicates the largest variation unexplained by the first PC in a direction orthogonal to the first PC, and so on. Fig. 4 shows how a 3-dimensional collinear data set can be represented in a reduced 2-dimensional space using only two principal components. The number of the retained PCs is usually less than the number of measured variables.

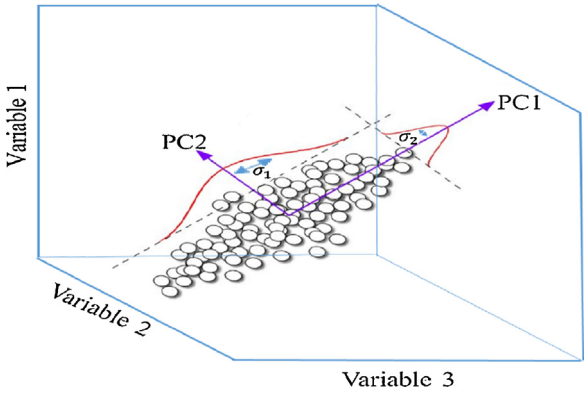

Fig. 4 – Principle of PCA.

A key step in the building of PCA model is to determine the number of PCs, l, that are required to adequately capture the major variability in the data sets. The goodness of the PCA model depends on a good choice of how many PCs are retained (Qin and Dunia, 2000). The first (l) largest principal components normally describe the most of the variance of the data. On the other hand, the smallest principal components are considered as a noise contributor. Too few components imply that there are not enough dimensions to represent the process variability, which degrades the prediction quality of the PCA model. While too many components imply that one can introduce noise and the model fails to capture some of the information. A number of techniques have been proposed to determine the number of PCs to be retained in a PCA model including cross validation (Li et al., 2002), Scree plot (Zhu and Ghodsi, 2006), and cumulative percent variance (CPV). In this study, the CPV technique will be used to determine the number of PCs for PCA model. The CPV is defined as follows: $CPV(l) = \frac{\sum_{i=1}^{l}\lambda_i}{trace(\Sigma)} \times 100$. Once the number of principal components l is determined, the PCA algorithm decomposes $\mathbf{X}_s$ into two orthogonal parts: an approximated data matrix $\widehat{\mathbf{X}}$ and a residual data matrix $\mathbf{E}$, i.e.,

$$\mathbf{X}_s = \sum_{i=1}^{l} t_i p_i^T + \sum_{i=l+1}^{m} t_i p_i^T = \widehat{\mathbf{X}} + \mathbf{E}. \tag{3}$$

Of course, if some of the variables in the data set are collinear or highly correlated, then a smaller number of principal components l are required to explain the majority of the variance in the data. In practice, the variance left unexplained by the PCs is captured by the residual subspace, which are often associated with the instrument or process noise.

4.2. PCA-based detection indices

As shown in Eq. (3), any measured vector $\mathbf{x}$ can be expressed using PCA as the sum of two orthogonal parts, approximated vector $\widehat{\mathbf{x}}$ and residual vector $\mathbf{e}$ (see Fig. 5), corresponding to the projection onto the PC subspace S_p and residual subspace S_r, respectively. In anomaly-detection using PCA, a PCA model is constructed using fault-free data, and then the model is used to detect faults using one of the detection indices, such as Hotelling's T^2 and Q statistics, which are described next.

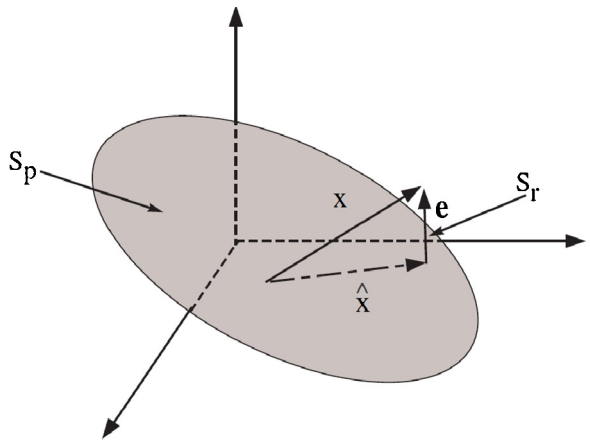

Fig. 5 – Geometric principle of PCA.

4.2.1. *Hotelling's T^2 statistic*

The T^2 statistic measures the variations in the principal components or score vectors at different time samples. T^2 at any instance of time is defined as follows (Hotelling, 1933):

$$T^2 = x^T \widehat{P} \widehat{\Lambda}^{-1} \widehat{P}^T x = \sum_{i=1}^{l} \frac{t_i^2}{\lambda_i}, \tag{4}$$

where the matrix $\widehat{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_l)$, is a diagonal matrix containing the eigenvalues associated with the l retained principal components. The threshold value used for the T^2 statistic can be computed as follows (Hotelling, 1933):

$$T^2_{l,n,\alpha} = \frac{l(n-1)}{n-l} F_{l,n-l,\alpha}, \tag{5}$$

where n is the number of samples in the data, l is the number of retained PCs, α is the level of significance (α usually takes values between 1% and 5%), and $F_{l,n-l}$ is the Fisher F distribution with l and $n-l$ degrees of freedom. When the number of observations, n, is rather large, the T^2 statistic threshold can be approximated with a χ^2 distribution with l degrees of freedom, i.e., $T^2_\alpha = \chi^2_{l,\alpha}$. These threshold values are computed using fault-free data. For new testing data, when the value of T^2 exceeds the value of the threshold, $T^2_{l,n,\alpha}$ or T^2_α, a fault is declared.

4.2.2. *Q statistic or squared prediction error (SPE)*

The Q statistic or Rao-statistic (also referred to as the squared prediction error, SPE) measures the projection of a data sample on the residual subspace, which provides an overall measure of how a data sample fits the PCA model. Q is defined as the sum of squares of the residuals obtained from the PCA model, i.e., (Qin, 2003):

$$Q = \mathbf{e}^T \mathbf{e}. \tag{6}$$

The upper control limit of this statistic is defined as (Jackson and Mudholkar, 1979):

$$Q_\alpha = \varphi_1 \left[\frac{h_0 c_\alpha \sqrt{2\varphi_2}}{\varphi_1} + 1 + \frac{\varphi_2 h_0 (h_0 - 1)}{\varphi_1^2} \right], \tag{7}$$

where c_α is the value of the normal distribution with α level of significance, $\varphi_i = \sum_{j=l+1}^{m} \lambda_j^i$ for $i = 1, 2, 3$, and $h_0 = 1 - \frac{2\varphi_1\varphi_3}{3\varphi_2^2}$. This value of threshold is calculated based on the assumptions that the measurements are time-independent and multivariate normally distributed. The Q fault detection index is very sensitive to modeling errors and its performance largely depends on the choice of the number of retained principal components, l, (Qin, 2003). The PCA fault detection algorithm is summarized next.

(1) **Given**:
- A training fault-free data set that represents the normal process operations and a testing data set (possibly faulty data).

(2) **Data preprocessing**
- Scale the data to zero mean and unit variance.

(3) **Build the PCA model using the training fault-free data**
- Compute the covariance matrix, Σ, using Eq. (2).
- Calculate the eigenvalues and eigenvectors of Σ and sort the eigenvalues in decreasing order.
- Determine how many principal components to be used. Many techniques can be used in this regards. In this work, the *CVP* criterion is used.
- Express the data matrix as a sum of approximate and residual matrices as shown in Eq. (3).
- Compute the control limits for the statistical model (e.g., the Q_α statistic limits).

(4) **Test the new data**
- Scale the new data.
- Generate a residual vector, **e**, using PCA.
- Compute the monitoring statistic (Q or T^2 statistics) for the new data using Eq. (4) or (6).

(5) **Check for anomalies**
- Declare an anomaly when new data exceeds the control limits (e.g., $Q \geq Q_\alpha$).

Unfortunately, the T^2 and Q statistics use only the observed data at the current time point alone for making decision about the process performance at the current time point. They take into account only the present information of the process thus they have a short memory. For this reason the T^2 and Q statistics are also called detection indices without memory. Consequently, these detection indices are relatively insensitive to small changes in the process variables, and thus may result in missed detections (Montgomery, 2005). These drawback of the T^2 and Q statistics motivate the use of other alternatives in order to surmount these disadvantages. Note that the ability to detect smaller parameter shifts can be improved by using a chart based on a statistic that corporate information from past samples in addition to current samples. In this study, anomaly detection technique which is based on PCA model and MEWMA control scheme will be developed in order to surmount these drawbacks and improve detection performance compared to the conventional PCA based anomaly detection method. A succinct introduction to the basic ideas behind MEWMA monitoring scheme is exposed in the subsequent section.

5. Multivariate EWMA statistical control scheme

Control charts are one of the most frequently used procedures in statistical process control (SPC), and have been widely used as a monitoring tool in quality engineering to detect the existence of possible anomalies in the mean or variance of process measurements. Many control charts are referenced in the bibliography, and they can be broadly

categorized into main classes: univariate and multivariate techniques (Montgomery, 2005; Bissell, 1994). The univariate control charts such as Shewhart, cumulative summation (CUSUM) (Page, 1954), and EMWA (Montgomery, 2005) have been designed to essentially to monitor only one process variable. However, modern industrial processes often present a large number of highly correlated process variables. This is the area where univariate control charts are unable to explain different aspects of the process and, therefore, it is not appropriate for modern day processes. Moreover, to monitor several different process variables in the same time multivariate statistical monitoring charts such as multivariate Shewhart (Montgomery, 2005), multivariate EWMA (MEWMA) (Lowry et al., 1992) and multivariate CUSUM (MCUSUM) (Montgomery, 2005) were developed in analogy with the univariate charts. In fact, most commonly used multivariate control charts are the natural extension of the univariate charts, e.g., Hotelling's T^2 charts (Hotelling, 1947), MEWMA charts and MCUSUM charts (Montgomery, 2005; Lowry et al., 1992). A multivariate SPC charts take into account the additional information due to the correlation between a process variables while univariate SPC charts do not. These concepts may be used to develop more efficient control charts than the simultaneous operation of several univariate control charts.

The MEWMA chart was first proposed by Lowry et al. (1992) to monitor mean shifts of a multivariate process. This is a multivariate extension of the univariate EWMA chart proposed by Roberts (1959). This monitoring chart is constructed based on a weighted moving average of all observed data and available at the current time point. The MEWMA is utilized when there are several correlated process variables to be monitored simultaneously where detecting faults with small magnitudes is of interest. Suppose that we observe $\mathbf{X}_t = (X_1, X_2, \ldots, X_m)^T$, a m-dimensional set of observations at time t. A MEWMA control chart is proposed by Lowry et al. (1992) as follows:

$$\mathbf{Z}_t = \mathbf{R}\mathbf{X}_t + (\mathbf{I}_{m\times m} - \mathbf{R})\mathbf{Z}_{t-1}, \tag{8}$$

where $\mathbf{R} = diag(r_1, r_2, \ldots, r_m)$ which is a diagonal matrix with $r_1, r_2, \ldots, r_m$ on the main diagonal, and m is the number of variables; $0 < r_j \leq 1$ is a weighting parameter for j-th component of $\mathbf{X}$, for $j = 1, 2, \ldots, m$, $\mathbf{I}_{m\times m}$ is the identity matrix, Z_i is the ith EWMA vector, and X_i is the ith observation vector $i = 1, 2, \ldots, n$. The initial value Z_0 is usually obtained as equal to the in-control mean vector of the process. Generally, in quality control, a smaller value of r leads to quicker detection of smaller shifts (Lucas and Saccucci, 1990). Indeed, r should be adjusted to a value appropriate for the characteristic of the monitored process. Usually, the larger the shift is, the greater the r is. The value of r is usually set between 0.2 and 0.3 (Hunter, 1986). It can be noticed that if $\mathbf{R} = \mathbf{I}$, then the MEWMA control chart is equivalent to the T^2 chart. In this case, a MEWMA chart has been automatically changed into T^2 chart.

In practice, if there is no priori reason to weight different components differently, then we can simply choose $r_1 = r_2 = \cdots = r_m = r$. In this case Eq. (8) can be written as follows:

$$\mathbf{Z}_t = r\mathbf{X}_t + (1 - r)\mathbf{Z}_{t-1}. \tag{9}$$

The MEWMA decision function, $\mathbf{V}_t^2$, can be calculated recursively as follows (Lowry et al., 1992):

$$\mathbf{V}_t^2 = \mathbf{Z}_i^T \boldsymbol{\Sigma}_{\mathbf{Z}_t}^{-1} \mathbf{Z}_t, \tag{10}$$

where $\boldsymbol{\Sigma}_{\mathbf{Z}_t}$ is the variance–covariance matrix of $\mathbf{Z}_t$. When $r_1 = r_2 = \cdots = r_p = r$, the variance–covariance matrix of $\mathbf{Z}_t$ can be simplified to:

$$\boldsymbol{\Sigma}_{\mathbf{Z}_t} = \frac{r}{(2-r)}[1 - (1-r)^{2n}]\boldsymbol{\Sigma}, \tag{11}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the input data. The MEWMA chart statistic is usually constructed in terms of the asymptotic covariance matrix. When t becomes large, the covariance matrix converges to: $\boldsymbol{\Sigma}_{Z_i} = \left(\frac{r}{(2-r)}\right)\boldsymbol{\Sigma}$.

Under nominal conditions, the statistic Z is distributed according to the Gaussian law with zero mean and variance–covariance matrix $\boldsymbol{\Sigma}_{Z_i}$, $Z \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{Z_i})$. The distribution of the statistic Z in the presence of additive mean shift μ_1 is given as: $Z \sim \mathcal{N}(r\sum_{j=1}^{n}[(1-r)^{n-j}\theta], \boldsymbol{\Sigma}_{Z_i})$. The MEWMA chart declares the presence of anomaly when $\mathbf{V}_t^2 > h$, where h is the control limit. The distribution of $\mathbf{V}_t^2$ under in-control condition is χ_p^2. However, because the variables in the time series $\mathbf{V}_t^2$, $t = 1, 2, \ldots$ are correlated, the control limit h cannot simply be chosen to be $(1-\alpha$-th) quantile $\chi_{1-\alpha,p}^2$ of the χ_P^2 distribution. One of the main troubles on this chart is the selection of the h. The value of h can be calculated by simulation to achieve a specific control limits. Various authors have used theoretical derivation, Markov chain approximation, integral equation approximation, and Monte Carlo simulation, or combinations of the three techniques to compute the control limit h according to the parameters r, p, and α (Runger and Prabhu, 1996; Rigdon, 1995). Bodden and Rigdon (1999) proposed an algorithm to find the control limit h in order to respect a given number of false alarm and a given r.

6. Anomaly detection using a PCA-based MEWMA control scheme

In this section, PCA is integrated with MEWMA to develop a new anomaly detection scheme with a higher sensitivity to small or moderate anomalies in the data. Toward this end, PCA is used to represent a matrix of the process measurements as the sum of two orthogonal parts (an approximated data matrix and a residual data matrix) as shown in Eq. (3). In PCA model, the principal components associated with large eigenvalues capture most of the variations in the data, where, ones associated with small eigenvalues mostly represent noise and are sensitive to the observations that are inconsistent with the correlation among the variables (Jobson, 1992; Donnell et al., 1994). Therefore, the smallest principal components (i.e., associated with small eigenvalues) should be useful in anomaly detection. The smallest ignored PCs can be used as an indicator about the existence or absence of faults. When the monitored process is under healthy conditions (no anomaly), the least important principal components are close to zero. However, when a anomaly occurs, then they tend to largely deviate from zero indicating the presence of a new condition that is significantly distinguishable from the normal healthy mode. In this paper, MEWMA is used to enhance process monitoring through its integration with PCA. Because of the ability of the MEWMA control scheme to detect small/moderate changes in the data, this technique

Table 1 – PCA-based MEWMA fault detection algorithm.

Step	Action
1.	**Given**: • A training fault-free data set that represents the normal process operations and a testing data set (possibly faulty data). • The parameters of the MEWMA control scheme: smoothing parameter r and the probability of false alarm α.
2.	**Data preprocessing** • Scale the data to zero mean and unit variance.
3.	**Build the PCA model using the training fault-free data** • Express the data matrix as a sum of approximate and residual matrices as shown in Eq. (3). • Compute the ignored principal components $\widetilde{\mathbf{t}}^j$, using PCA. • Compute the MEWMA control limits.
4.	**Test the new data** • Scale the new data. • Compute the principal components $\widetilde{\mathbf{t}}^j$, using PCA. • Compute the MEWMA decision function, V_t^2.
5.	**Check for anomalies** • Declare a fault when the MEWMA decision function, V_t^2, exceeds the control limits.

is appropriate to improve the detection of moderate anomalies. Thus, this work exploits the advantages of the MEWMA control scheme to improve anomaly detection over the conventional PCA-based methods. Toward this end, the MEWMA control scheme is used to monitor the ignored principal components, which correspond to the small eigenvalues of the PCA model.

6.1. PCA-based MEWMA process monitoring algorithm

In this approach, the MEWMA monitoring scheme is applied using the principal components ignored (which have smallest variances) from the PCA model. If the matrix of ignored principal components is defined as $\widetilde{\mathbf{T}} = [\mathbf{t}^{l+1}, \ldots, \mathbf{t}^j, \ldots \mathbf{t}^m]$, where $\mathbf{t}^j \in R^n$, i.e., $\mathbf{t}^j = [\mathbf{t}_1^j, \ldots, \mathbf{t}_t^j, \ldots, \mathbf{t}_n^j]$, then the MEWMA function can be computed using the residuals of the jth principal component as follows:

$$z_t^j = r\mathbf{t}_t^j + (1-r)z_{t-1}^j, \quad j \in [1, m-l]. \tag{12}$$

The MEWMA decision function, $\mathbf{V}_t^2$, can be calculated recursively as follows (Lowry et al., 1992):

$$\mathbf{V}_t^2 = \mathbf{Z}_i^T \boldsymbol{\Sigma}_{\mathbf{Z}_t}^{-1} \mathbf{Z}_t, \tag{13}$$

where $\boldsymbol{\Sigma}_{\mathbf{Z}_t}$ is the variance–covariance matrix of $\mathbf{Z}_t$.

In this case, since the MEWMA control scheme is applied on the ignored $m-l$ principal components, one MEWMA decision function will be computed to monitor the process. However, this approach can only detect the presence of anomalies, i.e., it cannot determine their locations. This approach is summarized in Table 1.

In the next section, the performance of the proposed PCA-based MEWMA fault detection method will be evaluated and compared to that of the conventional PCA anomaly detection scheme through their application to monitor would rotor induction machines.

7. Results and discussion

In this section, the proposed PCA-based MEWMA anomaly detection scheme is applied in order to detect abnormalities in ozone measurements caused by air pollution or any incoherence between the different network sensors or sensor faults in the framework of regional ozone surveillance network in Upper Normandy. The performance of the proposed method is compared to that obtained with the conventional PCA approach and to that declared by Air Normand air monitoring association.

7.1. Problem setting

In this study, the data that we use were extracted from the Upper Normandy region. The ozone concentrations data are measured each 15 min in order to limit spatial and temporal sampling problems. The data series of ozone concentrations measured from 11 August to 19 August, 2006 with a total number of 773 observations were used to develop a PCA model without faults. Plots of the original ozone concentration times series and of the corresponding auto-correlation functions (ACF) are shown in Fig. 6. Only the curves of the three stations 'SRC', 'QUI' and 'ND2' are plotted for better readability of the figures. These three stations behave like the others network stations.

From Fig. 6, the ACF graphics shows an apparent periodicity of 24 h. It is well known that the distance between extremum points in the autocorrelation functions gives the period of the time series. We suspect that this periodicity is related to the diurnal cycle of ozone which is primarily caused by the diurnal temperature cycle. This periodic variation is due to the cycle of solar radiation (day/night) which is closely related to the mechanism of formation of this pollutant. We also can see the similarity between the autocorrelation functions of ozone concentrations of the majority of network stations. Monitoring such data therefore requires an initial processing step where such explainable patterns and seasonality are removed. PCA can handle the high dimension of the measurement network and the high degree of correlation among some variables. The purpose is to detect abnormalities in ozone measurements.

7.2. PCA modeling

Firstly, a PCA model is build using training data set. The fault-free data used to develop the model was arranged in a matrix $\mathbf{X}$ with 773 rows (samples) and 7 columns (ozone concentration variables). These data matrix are scaled (to be zero mean with a unit variance), and then used to construct a PCA model.

The scaled fault-free data matrix is used to construct a PCA model, and the computed principal components are shown in Fig. 7. Indeed, the principal components (PCs) are linear combinations of the original ones and are uncorrelated. Although PCs represent directions (or patterns) that explain most of the observed variability, their interpretation is, however, not always simple. More specifically, they are just mathematical constructs chosen to represent the variance as efficiently as possible and to be orthogonal to each other. It can be noticed from Fig. 7 that the principal components $t3, \ldots, t7$ represent mainly noise while the first two principal components $t1$ and $t2$ capture most of the important variations in the data. More

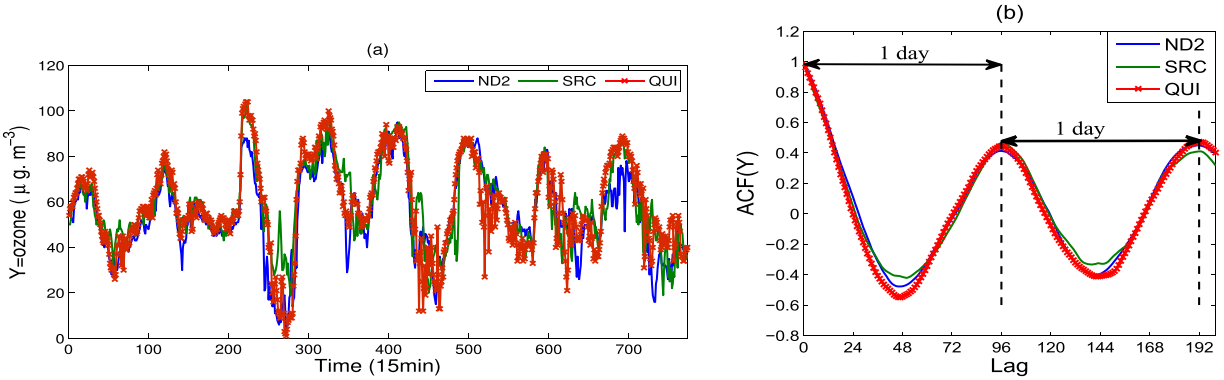

Fig. 6 – (a) Quarter-hourly ozone time series and (b) ACF of ozone time series.

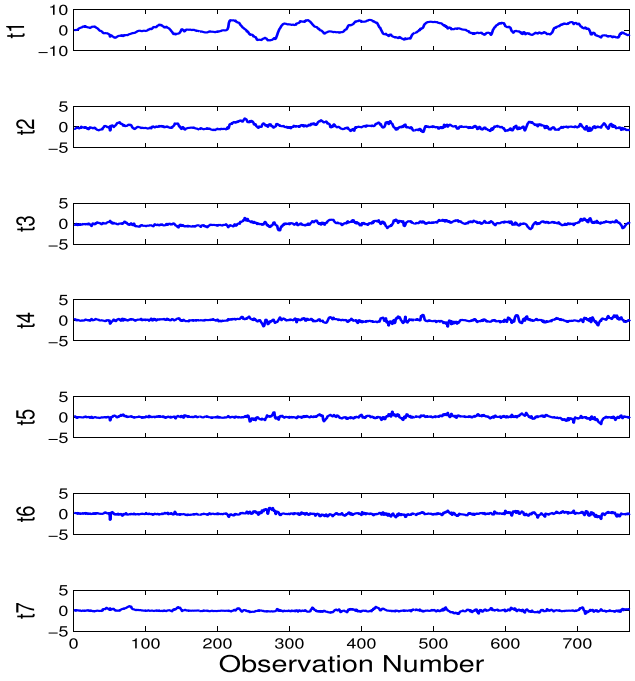

Fig. 7 – The principal components of the fault-free data.

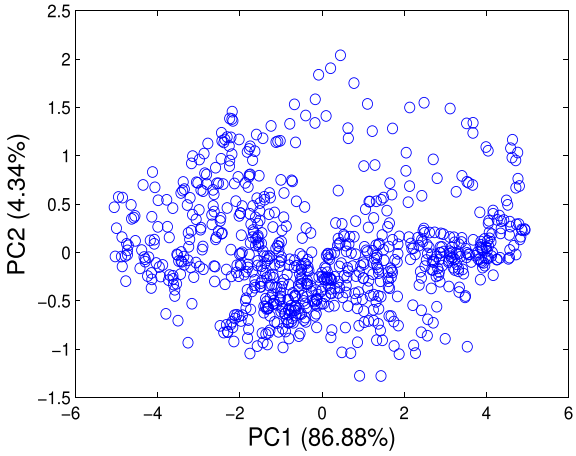

Fig. 8 – PC1 versus PC2.

specifically, the first principal component, $t1$, is the direction of greatest variability in the data (capture 86%:88% of the total variations in the data). The second, $t2$, is the next orthogonal (uncorrelated) direction of greatest variability (capture 4%:34% of the total variations in the data). In this case study, $t1$ and $t2$ capture most of the important variations in the data.

In PCA, most of the important variations in the data are usually captured in few principal components corresponding to the largest eigenvalues. In this work, the cumulative percent variance (CPV) method is used to determine the optimum number of retained principal components. Using a CPV threshold value of 90%, only the first two principal components will be retained since they capture 86.88% and 4.34% of the total variations in the data.

Indeed, the principal components are linear combination of the original ones, and are uncorrelated with one another. To determine whether principal components are uncorrelated, the scatter plot of PC1 and PC2 is examined. If there were a noticeable relationship in this plot, it would be attributed to non-linear relationships in the data. The PC technique removes all linear correlations and results in a scatter plot when the non-linear relationships are small or nonexistent. Fig. 8 shows the bivariate scores plot of PC1 versus PC2 and shows that PC1 and PC2 are uncorrelated. The PCA technique removes all linear correlations.

Fig. 9 presents standardized measurements and estimation for the whole measurements network, the estimations being given by the PCA model. By taking into account the nature of considered process, the results are very satisfactory. With this PCA model based on the first two PCs, the ozone concentrations is generally correctly estimated. However, for some variables we can have modeling errors as shown in Fig. 9 (stations ND2, TAN and QUI). In conclusion, the linear PCA was able to model the relations between the various variables. However as we could not it, certain variables being less better estimated than others, we now will examine the effect of the modeling errors on the fault detection phase.

7.3. *Detection results*

In this section, the anomaly detection abilities of the developed PCA-based MEWMA anomaly detection approach will be assessed using the Upper Normandy ozone data which are completely independent from the training data used to construct the reference PCA model. To evaluate the performance of the developed method, the detection results of the proposed method are compared to that declared by Air Normand, and to that of conventional PCA. Three different testing data sets have been used to evaluate the performance of the PCA-based MEWMA anomaly detection scheme. The first sample covers the period from 11 June 2006 to 09 July 2006, a period of 27 days. The second sample covers the period from 19 August 2006 to 8 September 2006, a period of 21 days. The latter covers the period from 9 September 2006 to 10 October is a

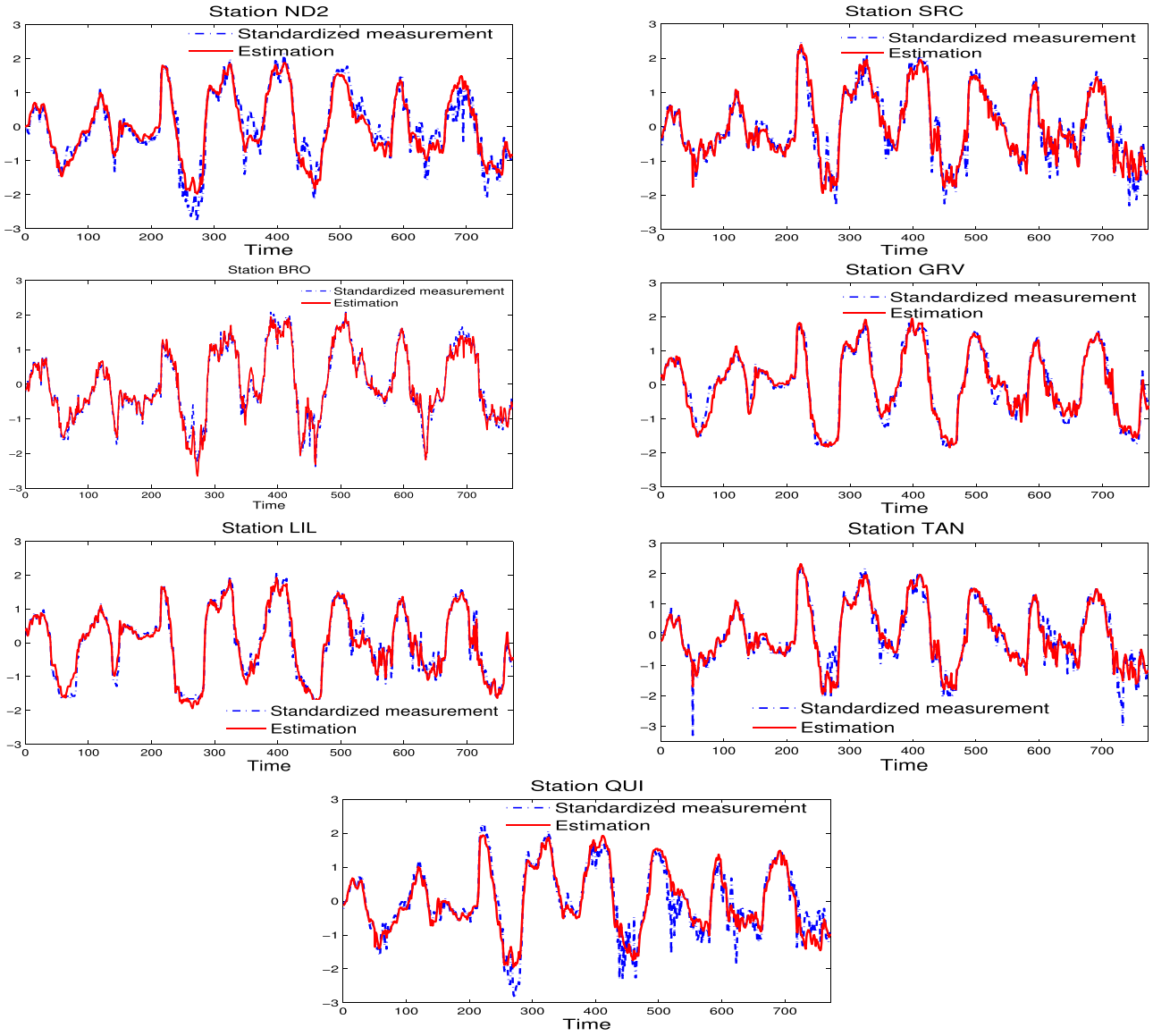

Fig. 9 – Measurements and estimation of ozone level for the three station.

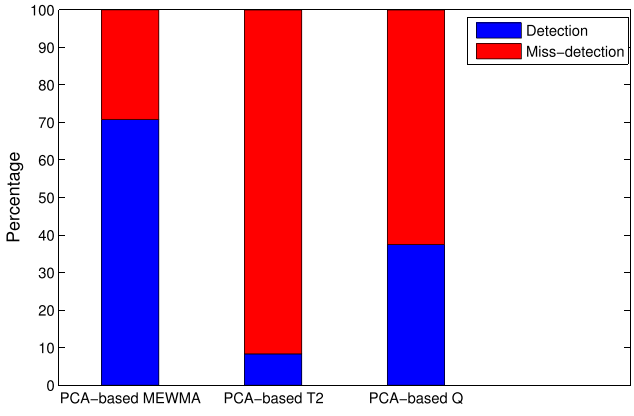

Fig. 10 – Compare detection results.

period of 29 days. When the developed PCA-based MEWMA anomaly detection scheme is applied using the fault-free data, the MEWMA threshold value is found to be $h(\alpha) = 9.65$ for a smoothing parameter $r = 0.25$ and a false alarm probability of $\alpha = 0.005$. The detection results are given in Table 2 and are visually illustrated in Fig. 10.

In Table 2, the first seven columns present the results of analysis given by Air Normand experts. The first column presents the date of an anomaly observed by experts of Air Normand. The second and third columns present the time and the maximum peak intensity. The column 4 presents the station name where the anomaly has occurred and columns 5, 6 and 7 show the beginning, the end and the duration of this anomaly. The column 8 shows the results of detection given by PCA-based MEWMA anomaly detection scheme. The columns 9 and 10 show the results of detection given by the conventional PCA detection indices, T^2, and Q, respectively. If the result is *yes*, then it is a correct detection. If the result is *no*, then it is a missed detection. For example take the first two lines to describe how to read this table. The first line indicates that the station 'LIL' has measured abnormal level ozone 12/06/2006 between 11:30 and 12:45 for a total duration of 0:45 min and the anomaly peak has occurred at 11:45 with a maximum intensity level in 141.3 μg. The developed PCA-based MEWMA anomaly detection scheme does not detect this anomaly (see column 8 detection). The results of the T^2 and Q statistics shown in columns 9 and 10, respectively, show that the conventional PLS was unable to detect this anomaly. In the second line, ND2 and LIL stations have presented abnormalities on 13/06/2006.

Table 2 – Detection results.

Date	Air Normand detection						PCA-MEWMA	PCA-T^2	PCA-Q
	Hour	Intensity	Places	Beginning	End	Duration			
12/06/2006	11:45	141	LIL	11:30	12:15	0:45	No	No	No
13/06/2006	13:15	168	LIL	12:30	13:45	1:15	Yes	No	Yes
		181	ND2	12:15	14:15	2:00	Yes	No	No
17/06/2006	08:00	132	SRC	7:15	10:15	3:00	No	No	No
	08:30	141	TAN	8:00	9:00	1:00	Yes	No	No
23/06/2006	14:15	137	LIL	13:00	15:00	2:00	No	No	No
	14:30	126	ND2	13:00	15:15	2:15	No	No	No
	14:45	127	QUI	13:15	15:15	2:00	No	No	No
30/06/2006	08:00	144	TAN	7:15	8:15	1:00	Yes	No	No
03/07/2006	08:15	244	TAN	8:15	9:15	1:00	Yes	Yes	Yes
	10:15	242	TAN	9:15	11:15	2:00	Yes	Yes	Yes
	9:30	179	LIL	9:00	10:15	1:15	Yes	No	Yes
	10:00	166	QUI	9:15	10:15	1:00	Yes	No	No
04/07/2006	07:45	201	ND2	6:30	10:00	3:00	Yes	No	Yes
05/09/2006	09:45	180	LIL	7:45	10:45	3:00	Yes	No	No
	09:45	115	TAN	8:15	11:00	2:45	No	No	No
06/09/2006	09:45	182	LIL	8:15	10:30	2:15	Yes	No	Yes
	11:15	168	LIL	10:30	13:15	2:45	Yes	No	No
	14:00	168	ND2	13:15	15:00	1:45	Yes	No	No
	14:30	168	GRV	13:45	15:00	1:15	Yes	No	No
10/09/2006	09:30	167	QUI	7:30	10:00	2:30	Yes	No	No
	09:45	146	LIL	8:45	10:30	1:45	Yes	No	No
	11:00	180	TAN	10:15	11:30	1:15	Yes	No	No
	12:00	166	GRV	11:30	12:45	1:15	No	No	No

The PCA-based MEWMA scheme has correctly detected these anomalies. The results using the Q statistic given in column 10 show that it could successfully detect this anomaly. However, Hotelling's T^2 statistic was unable to detect this anomaly. This result may be explained by the fact that the T^2 statistic provides a measure of the deviation in the PCs that are of greatest importance to the normal process condition. Thus, the normal operating region defined by the T^2 control limits is usually larger than that defined by the Q control limits. Therefore, anomalies with moderate magnitudes can easily exceed the Q threshold, but not the T^2 threshold, which makes the Q statistic usually more sensitive than T^2 for this anomaly. By comparing the results obtained by the PCA-based MEWMA detector and results declared by Air Normand, we note that the PCA-based MEWMA detector has detected almost the totality of anomalies (see Table 2 and Fig. 10). For our application, the proposed fault anomaly algorithm improves the anomaly detection compared to classical detection indices Q and T^2. The developed PCA-based MEWMA anomaly detection algorithm takes very little time to give its verdict. Hence, the proposed algorithm can be used as an automatic tool of abnormal ozone peaks (or sensors faults) detection in the framework of regional air quality monitoring networks.

8. Conclusion

In this paper, an anomaly detection scheme based on principal component analysis is proposed to monitor the ozone concentrations in the Upper Normandy region, France. To enhance anomaly detection a new PCA-based monitoring strategy combining PCA with the multivariate exponentially weighted moving average (MEWMA) monitoring scheme is proposed. In the proposed approach, MEWMA control scheme is applied on the ignored principal components (which have smallest variances) to detect the presence of anomalies. The proposed PCA-based MEWMA anomaly detection scheme is successfully applied to data of the ozone concentrations collected from the Upper Normandy region, France. For this application, the PCA-based MEWMA scheme improves the anomaly detection compared to that or the conventional PCA-based monitoring charts. The results indicate that the PCA-based MEWMA test can be used as an automatic tool to detect abnormal ozone measurements.

References

Banimostafa, A., Papadokonstantakis, S., Hungerbühler, K., 2012. Evaluation of EHS hazard and sustainability metrics during early process design stages using principal component analysis. Process saf. Environ. Prot. 90 (1), 8–26.

Bissell, D., 1994. Statistical Methods for SPC and TQM, vol. 26. CRC Press.

Bodden, K.M., Rigdon, S.E., 1999. A program for approximating the in-control ARL for the MEWMA chart. J. Qual. Technol. 31 (1), 120–123.

Box, G.E.P., 1954. Some theorems on quadratic forms applied in the study of analysis of variance problems: effect of inequality of variance in one-way classification. Ann. Math. Stat. 25, 290–302.

Brankov, E., Henry, R., Civerolo, K., Hao, W., Rao, S., Misra, P., Bloxam, R., Reid, N., 2003. Assessing the effects of transboundary ozone pollution between Ontario, Canada and New York, USA. Environ. Pollut. 123 (3), 403–411.

Brulfert, G., Galvez, O., Yang, F., Sloan, J., 2007. A regional modelling study of the high ozone episode of June 2001 in southern Ontario. Atmos. Environ. 41, 3777–3788.

Chen, Q., Kruger, U., Meronk, M., Leung, A., 2004. Synthesis of T2 and Q statistics for process monitoring. Control Eng. Pract. 126, 745–755.

Chen, W., Tang, H., Zhao, H., 2015. Diurnal, weekly and monthly spatial variations of air pollutants and air quality of Beijing. Atmos. Environ. 119, 21–34.
Chiogna, M., Pauli, F., 2011. Modelling short-term effects of ozone on morbidity: an application to the city of Milano, Italy, 1995–2003. Environ. Ecol. Stat. 18 (1), 169–184.
Detournay, A., Meur, S.L., Delmas, V., 2007. Understanding of the atypical ozone peaks phenomenon observed around the petrochemical industrial zone of Port-Jerome in Upper Normandy, France. Pollut. Atmos. 196, 405–422.
Donnell, D., Buja, A., Stuetzle, W., 1994. Analysis of additive dependencies and concurvities using smallest additive principal components. Ann. Stat., 1635–1668.
Due nas, C., Fernández, M., Ca nete, S., Carretero, J., Liger, E., 2004. Analyses of ozone in urban and rural sites in Málaga (Spain). Chemosphere 56, 631–639.
Geladi, P., Kowalski, B., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 185, 1–17.
George, J., Chen, Z., Shaw, P., 2009. Fault detection of drinking water treatment process using PCA and Hotelling's T^2 chart. World Acad. Sci. Eng. Technol. 50, 970–975.
Harkat, M., Mourot, G., Ragot, J., 2006. An improved PCA scheme for sensor FDI: application to an air quality monitoring network. J. Process Control 16 (6), 625–634.
Harrou, F., Kadri, F., Chaabane, S., Tahon, C., Sun, Y., 2015. Improved principal component analysis for anomaly detection: application to an emergency department. Comput. Ind. Eng. 88, 63–77.
Herve, A., Lynne, J., 2010. Principal component analysis. Wiley Interdiscip. Rev.: Comput. Stat. 2, 433–459.
Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441.
Hotelling, H., 1947. Multivariate quality control illustrated by the air testing of sample bomb sights. In: Techniques of Statistical Analysis. McGraw-Hill, New York (Chapter II).
Hunter, J.S., 1986. The exponentially weighted moving average. J. Qual. Technol. 18 (4), 203–210.
Jackson, J., Mudholkar, G., 1979. Control procedures for residuals associated with principal component analysis. Technometrics 21, 341–349.
Jobson, J., 1992. Applied Multivariate Data Analysis, vol. 2. Springer Heidelberg.
Khan, F., Rathnayaka, S., Ahmed, S., 2015. Methods and models in process safety and risk management: Past, present and future. Process Saf. Environ. Prot. 98, 116–147.
Kourti, T., MacGregor, J., 1995. Process analysis, monitoring and diagnosis using multivariate projection methods: a tutorial. Chemom. Intell. Lab. Syst. 28 (3), 3–21.
Kourti, T., MacGregor, J., 1996. Multivariate SPC methods for process and product monitoring. J. Qual. Technol. 28 (4).
Li, B., Morris, J., Martin, E., 2002. Model selection for partial least squares regression. Chemom. Intell. Lab. Syst. 64 (1), 79–89.
Lowry, C.A., Woodall, W.H., Champ, C.W., Rigdon, S.E., 1992. A multivariate exponentially weighted moving average control chart. Technometrics 34 (1), 46–53.
Lucas, J., Saccucci, M., 1990. Exponentially weighted moving average control schemes: properties and enhancements. Technometrics 32 (1), 1–12.
MacGregor, J., Kourti, T., 1995. Statistical process control of multivariate processes. Control Eng. Pract. 3 (3).
Mastrangelo, C., Runger, G., Montgomery, D., 1996. Statistical process monitoring with principal components. Qual. Reliab. Eng. Int. 12 (3), 203–210.
Montgomery, D.C., 2005. Introduction to Statistical Quality Control. John Wiley& Sons, New York.
Moshammer, H., 2010. Communicating health impact of air pollution. In: Villanyi, V. (Ed.), Air Pollution. InTech, ISBN 978-953-307-143-5, http://dx.doi.org/10.5772/10042, Available from: http://www.intechopen.com/books/air-pollution/communicating-health-impact-of-air-pollution.
Nawahda, A., 2016. An assessment of adding value of traffic information and other attributes as part of its classifiers in a data mining tool set for predicting surface ozone levels. Process Saf. Environ. Prot. 99, 149–158.
Page, E.S., 1954. Continuous inspection schemes. Biometrika, 100–115.
Patton, R.J., Chen, J., 1991. A review of parity space approaches to fault diagnosis. In: Proceedings of SAFEPROCESS'91, pp. 239–255.
Proyou, A., Toupance, G., Perros, P., 1991. A two year study of ozone behaviour at rural and forested sites in eastern France. Atmos. Environ. 25A (10), 2145–2153.
Qin, S., Dunia, R., 2000. Determining the number of principal components for best reconstruction. J. Process Control 10 (2), 245–250.
Qin, S., 2003. Statistical process monitoring: basics and beyond. J. Chemom. 17 (8/9), 480–502.
Qin, S., 2012. Survey on data-driven industrial process monitoring and diagnosis. Annu. Rev. Control 36 (2), 220–234.
Ralston, P., DePuy, G., Graham, J., 2001. Computer-based monitoring and fault diagnosis: a chemical process case study. ISA Trans. 40 (1).
Ranger, G., Alt, F., 1996. Choosing principal components for multivariate statistical process control. Commun. Stat. Theory Methods 25 (5), 909–922.
Rigdon, S., 1995. An integral equation for the in-control average run length of a multivariate exponentially weighted moving average control chart. J. Stat. Comput. Simul. 52 (4), 351–365.
Roberts, S.W., 1959. Control chart tests based on geometric moving averages. Technometrics 1 (3), 239–250.
Romagnoli, J., Palazoglu, A., 2006. Introduction to Process Control. CRC Press, United States of America.
Runger, G., Prabhu, S., 1996. A Markov chain model for the multivariate exponentially weighted moving averages control chart. J. Am. Stat. Assoc. 91 (436), 1701–1706.
Seinfeld, J., Pandis, S., 2006. Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. John Wiley & Sons, Inc., New York.
Sillman, S., 2003. Tropospheric ozone and photochemical smog. In: Sherwood Lollar, B. (Ed.), In: Treatise on Geochemistry, Environmental Geochemistry, vol. 9. Elsevier, pp. 407–431 (Chapter 11).
Vlachokostas, C., Nastis, S., Achillas, C., Kalogeropoulos, K., Karmiris, I., Moussiopoulos, N., Chourdakis, E., Banias, G., Limperi, N., 2010. Economic damages of ozone air pollution to crops using combined air quality and GIS modelling. Atmos. Environ. 44, 3352–3361.
Yin, S., Wang, G., Karimi, H., 2014. Data-driven design of robust fault detection system for wind turbines. Mechatronics 24 (4), 298–306.
Yin, S., Ding, S., Xie, X., Luo, H., 2014. A review on basic data-driven approaches for industrial process monitoring. IEEE Trans. Ind. Electron. 61 (11), 6418–6428.
Zdanevitch, I., 2001. Etude d'épisodes inexpliqués d'ozone, Rapport LCSQA, convension 41/2000. INERIS, Paris.
Zhu, M., Ghodsi, A., 2006. Automatic dimensionality selection from the scree plot via the use of profile likelihood. Comput. Stat. Data Anal. 51, 918–930.